PVStereo: Pyramid Voting Module for End-to-End Self-Supervised Stereo Matching

Hengli Wang, Rui Fan, Peide Cai and Ming Liu, *Senior Member, IEEE*

***Abstract*—Supervised learning with deep convolutional neural networks (DCNNs) has seen huge adoption in stereo matching. However, the acquisition of large-scale datasets with well-labeled ground truth is cumbersome and labor-intensive, making supervised learning-based approaches often hard to implement in practice. To overcome this drawback, we propose a robust and effective self-supervised stereo matching approach, consisting of a pyramid voting module (PVM) and a novel DCNN architecture, referred to as OptStereo. Specifically, our OptStereo first builds multi-scale cost volumes, and then adopts a recurrent unit to iteratively update disparity estimations at high resolution; while our PVM can generate reliable semi-dense disparity images, which can be employed to supervise OptStereo training. Furthermore, we publish the HKUST-Drive dataset, a large-scale synthetic stereo dataset, collected under different illumination and weather conditions for research purposes. Extensive experimental results demonstrate the effectiveness and efficiency of our self-supervised stereo matching approach on the KITTI Stereo benchmarks and our HKUST-Drive dataset. PVStereo, our best-performing implementation, greatly outperforms all other state-of-the-art self-supervised stereo matching approaches. Our project page is available at sites.google.com/view/pvstereo.**

***Index Terms*—Computer vision for automation, data sets for robotic vision, deep learning for visual perception.**

I. Introduction

HUMANS live in a three-dimensional (3D) world, but our eyes can only perceive objects in two dimensions. The miracle of human depth perception is due to our brain's ability to analyze the difference between the two two-dimensional (2D) images which are projected on the retinas of our eyes. In a broad sense, each pair of corresponding points on the retinas send signals to the binocular neurons in the primary visual cortex, which then estimates the relative positional difference between each pair of correspondence points [1]. This relative positional difference is generally referred to as *disparity* [2].

Manuscript received October 15, 2020; revised January 20, 2021; Accepted February 24, 2021. This paper was recommended for publication by Editor Cesar Cadena Lerma upon evaluation of the Associate Editor and reviewers' comments. This work was supported in part by the National Natural Science Foundation of China under grant No. U1713211, in part by the Collaborative Research Fund by Research Grants Council Hong Kong under Project No. C4063-18G, and in part by the HKUST-SJTU Joint Research Collaboration Fund under project SJTU20EG03, awarded to Prof. Ming Liu. *(Corresponding author: Ming Liu.)*

H. Wang, P. Cai and M. Liu are with the Department of Electronic and Computer Engineering, the Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR, China (email: hwangdf@connect.ust.hk; pcaiaa@connect.ust.hk; eelium@ust.hk).

R. Fan is with both the Department of Computer Science and Engineering and the Department of Ophthalmology, the University of California San Diego, La Jolla, CA 92093, U.S. (email: rui.fan@ieee.org).

H. Wang and R. Fan contributed equally to this work.

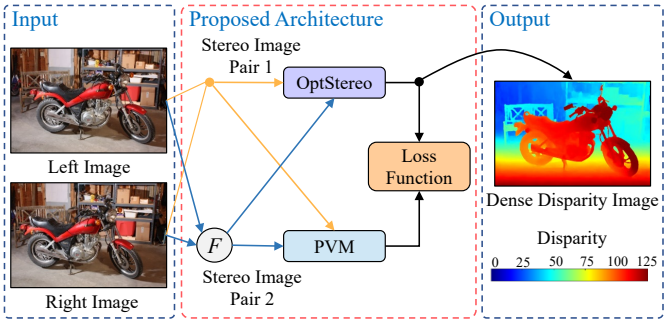

Fig. 1. A schema of our proposed self-supervised stereo matching approach, where OptStereo is combined with PVM for self-supervised disparity estimation, and F refers to a left-to-right flipping operation.

Similarly, two synchronized digital cameras can be utilized to extrapolate the 3D information of a given scenario. This process is typically known as *stereo vision* or *stereo matching* [2]. Stereo vision is a critical technology employed in many robotics and computer vision applications, such as freespace segmentation [3]–[5] and anomaly detection [6]–[8]. Existing stereo matching approaches are either mathematical modeling-based or data-driven ones. The former ones generally formulate stereo matching as block matching or energy minimization problems [9], while the latter ones typically employ data-driven classification and/or regression models, *e.g.*, convolutional neural networks (CNNs), to learn a feasible solution to stereo matching. With recent advances in deep learning, many researchers have resorted to deep CNNs (DCNNs) for stereo matching [10]–[12]. However, these approaches generally require a large amount of human-annotated training data to learn the best DCNN parameters. Such a data labeling process can be extremely time-consuming and labor-intensive. Furthermore, the limitation in DCNN generalization often fails these approaches when adapting to new scenarios in practice. Hence, there is a strong motivation to develop a self-supervised stereo matching approach, which does not require any human-annotated disparity ground truth to learn the best DCNN parameters [13].

Hence in this paper, we propose a novel approach for self-supervised stereo matching, as illustrated in Fig. 1. Specifically, we develop a module named *Pyramid Voting Module* (PVM), which can be deployed in any supervised stereo matching DCNN, converting it into a self-supervised approach. With the use of our PVM, researchers will no longer require any hand-labeled data to train DCNNs for dense stereo matching, which greatly alleviates the labor for disparity ground truth labeling. Moreover, existing stereo matching

DCNNs mainly rely on 3D convolutions [10]–[12] or coarse-to-fine paradigms [14]–[16]. Unfortunately, the former ones can consume a lot of computational resources, resulting in their limited ability to achieve real-time performance; while the latter ones generally suffer from the accumulated errors from the coarse pyramid level, which can in turn propagate to subsequent levels and further cause significant performance degradation. To address these issues, we propose a new DCNN architecture, referred to as *OptStereo*, which is motivated by traditional optimization-based approaches. Our OptStereo first builds multi-scale cost volumes, and then adopts a recurrent unit to iteratively update disparity estimations at high resolution. This novel architecture enables our OptStereo to 1) avoid the error accumulation problem in coarse-to-fine paradigms and 2) achieve a great trade-off between accuracy and efficiency. Furthermore, we publish the *HKUST-Drive dataset*, a large-scale synthetic stereo dataset created under different illumination and weather conditions, available at sites.google.com/view/pvstereo for research purposes. It contains 11568 pairs of stereo driving scene images and the corresponding dense ground-truth disparity images. To validate the effectiveness and efficiency of our proposed self-supervised stereo matching approach, we conduct extensive experiments on the popular KITTI Stereo benchmarks [17], [18] as well as our HKUST-Drive dataset. Extensive experimental results demonstrate that our best-performing implementation, *PVStereo*, outperforms all other self-supervised stereo matching approaches. The major contributions of this paper can be summarized as follows:

- PVM, a novel module capable of generating reliable semi-dense disparity images that can be used for supervising DCNN training.
- OptStereo, a novel DCNN architecture that can achieve a great trade-off between accuracy and efficiency for stereo matching.
- HKUST-Drive, a large-scale synthetic stereo dataset collected under different illumination and weather conditions for research purposes.

The remainder of this paper is organized as follows: Section II introduces existing traditional and data-driven approaches for stereo matching. Then, Section III presents our proposed framework for self-supervised stereo matching. The experimental results are illustrated in Section IV. Finally, Section V summarizes the paper.

II. Related Work

A. Traditional Stereo Matching Approaches

Traditional stereo matching approaches can be classified into three main categories: 1) local, 2) global, and 3) semi-global [9]. Local algorithms simply select a group of image blocks from the target image and match them with a fixed image block selected from the reference image [19], [20]. The desirable disparities possess either the lowest matching costs, *e.g.*, sum of absolute differences (SAD), or the highest correlation costs, *e.g.*, normalized cross-correlation (NCC) [21]. The optimization strategy utilized in local algorithms is typically referred to as winner-takes-all (WTA) [22].

Unlike local algorithms, global algorithms typically formulate stereo matching as an energy minimization problem, which can be solved by some Markov random field (MRF)-based optimization approaches, such as graph cuts (GC) [23] and belief propagation (BP) [24]. Semi-global matching (SGM) [25] approximates the MRF inference by performing cost aggregation along all directions in the image to improve both the accuracy and efficiency for stereo matching [26]. However, traditional methods are either inaccurate (local algorithms) or computationally intensive (global algorithms). With recent advances in deep learning, data-driven approaches can achieve a great trade-off between accuracy and efficiency.

B. Data-Driven Stereo Matching Approaches

1) Supervised Stereo Matching Approaches: Supervised stereo matching approaches can be classified into three categories: 1) learning better feature correspondences, 2) learning better regularization, and 3) learning dense disparity images in an end-to-end way. The first category of approaches utilize the learned distinguishable features to compute stereo matching costs, and then apply traditional cost aggregation and regularization for disparity estimation [27]. The second category of approaches learn both regularization and cost aggregation, *e.g.*, the spatial-variant penalty-parameters in SGM [28].

Recently, researchers have turned their focuses towards the third category, due to its excellent performance on public benchmarks. Such end-to-end approaches generally rely on 3D convolutions [10]–[12] or coarse-to-fine paradigms [14]–[16]. Specifically, Chang *et al.* [10] proposed PSMNet, a pyramid stereo matching network consisting of spatial pyramid pooling and several 3D convolutional layers. GwcNet [11] and AcfNet [12] were developed based on PSMNet for further performance improvement. However, 3D convolutions can consume a lot of computational resources, making these approaches difficult to perform in practice. To improve the DCNN inference speed, some researchers have adopted coarse-to-fine paradigms to replace 3D convolutions. Specifically, Tankovich *et al.* [14] proposed HITNet, which generates disparity predictions hierarchically from 1/64 resolution to 1/4 resolution. Similarly, Wang *et al.* [15] and Yee *et al.* [16] also followed this paradigm to improve the DCNN inference speed. However, these approaches typically have limited capability to recover errors from coarse resolutions, which can lead to significant performance degradation. Unlike the above-mentioned prior works, our proposed OptStereo employs a recurrent unit to iteratively updates disparity estimations at high resolution, which helps achieve a great trade-off between accuracy and efficiency for stereo matching.

2) Unsupervised Stereo Matching Approaches: To reduce the labor for disparity ground truth labeling, many researchers have proposed unsupervised stereo matching approaches [29]–[31]. Specifically, Zhong *et al.* [29] proposed a self-supervised stereo matching approach, which uses a novel training loss to exploit the loop constraint in image warping process and handle the texture-less areas. Given coarse information about the scenes and the optical system, Tulyakov *et al.* [30] developed an approach to generate disparity estimations in a

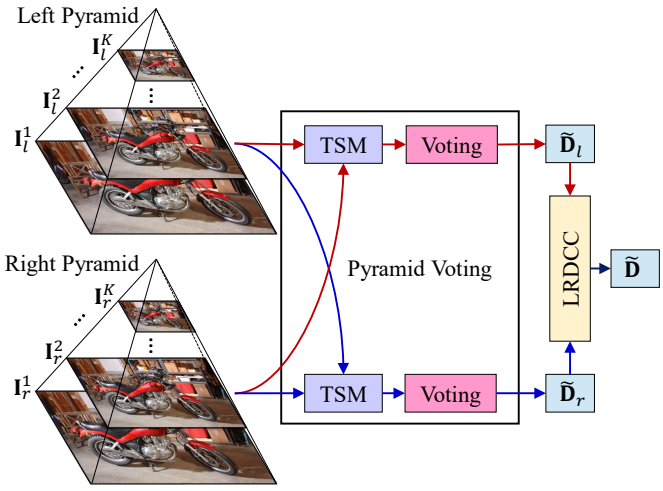

Fig. 2. An illustration of our proposed PVM. The input stereo images are processed to generate reliable semi-dense disparity images for supervising DCNN training. The red and blue paths are used to produce the left and right semi-dense disparity images, respectively.

weakly-supervised manner. Moreover, Liu *et al.* [31] proposed Flow2Stereo, which leverages the geometric constraints behind stereoscopic videos to perform disparity and optical flow estimation in a self-supervised manner. Different from these approaches, we propose PVM in this paper for reliable semi-dense disparity generation. The generated disparity images are then used to supervise DCNN training. Compared to the prior works, our PVM is easy-to-use, efficient and accurate. Extensive experimental results provided in Section IV demonstrate the superiority of our PVStereo over all other state-of-the-art self-supervised stereo matching approaches.

III. Methodology

A. Pyramid Voting Module

Our PVM can produce a reliable semi-dense disparity image $\tilde{\mathbf{D}}$ under a multi-scale disparity voting strategy, as illustrated in Fig. 2. The produced $\tilde{\mathbf{D}}$ can be utilized to supervise DCNNs in learning dense disparity estimation, and therefore, our PVM can convert any supervised stereo matching DCNN into a self-supervised approach.

Our PVM is designed based on two hypotheses:

1) confident disparities possess similar values, and
2) their matching costs or correlations are consistent,

regardless of image resolution. Therefore, our PVM aims at seeking out consistent disparities among multi-scale stereo image pairs from two pyramids. Given a pair of left and right stereo images $\mathbf{I}_l$ and $\mathbf{I}_r$, PVM first generates two groups of stereo image pairs, constructing a left and a right pyramid, respectively. One group is used to produce the left semi-dense disparity image $\tilde{\mathbf{D}}_l$ (see the red flow in Fig. 2), while the other one is used to produce the right semi-dense disparity image $\tilde{\mathbf{D}}_r$ (see the blue flow in Fig. 2). Each group contains a collection of K stereo image pairs at different scales, as illustrated as $(\mathbf{I}_l^1, \mathbf{I}_r^1), \ldots, (\mathbf{I}_l^K, \mathbf{I}_r^K)$. In this paper, $\mathbf{I}_{l,r}^k$ represents an $\mathbf{I}_{l,r}$ downsampled by a scale $k+\epsilon$, where $\epsilon \in (-1, 1)$ is a random scalar and $\epsilon = 0$ when $k = 1$. Each generated stereo image pair $(\mathbf{I}_l^k, \mathbf{I}_r^k)$ can separately produce a left and a right disparity image $\tilde{\mathbf{D}}_{l,r}^k$ via a traditional stereo matching algorithm (abbreviated as TSM in Fig. 2) [20]. Based on the above hypotheses, a representation $\mathbf{C}$ can be obtained:

$$\mathbf{C}(\mathbf{p}) = \left(\sqrt{\frac{1}{K}\sum_{k=1}^{K}(\tilde{\mathbf{D}}_{l,r}^k(\mathbf{p}) - \frac{1}{K}\sum_{k=1}^{K}\tilde{\mathbf{D}}_{l,r}^k(\mathbf{p}))^2}, \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\tilde{\mathbf{c}}_{l,r}^k(\mathbf{p}) - \frac{1}{K}\sum_{k=1}^{K}\tilde{\mathbf{c}}_{l,r}^k(\mathbf{p}))^2} \right), \tag{1}$$

where $\mathbf{p}$ is an image pixel; and $\tilde{\mathbf{c}}_{l,r}^k \in [0, 1]$ denotes the normalized inverse stereo matching cost or normalized correlation (a better stereo matching corresponds to a higher $\tilde{\mathbf{c}}_{l,r}^k$ value). A voting map $\mathbf{V}$ can then be obtained, where $\mathbf{V}(\mathbf{p}) = \delta(\mathbf{C}(\mathbf{p}, 1), \kappa_1) + \delta(\mathbf{C}(\mathbf{p}, 2), \kappa_2)$. κ_1 and κ_2 are two thresholds; and $\delta(x, y) = 0$ when $x < y$, otherwise, $\delta(x, y) = 1$. Generating a denser $\tilde{\mathbf{D}}_{l,r}$ requires higher κ_1 and κ_2. By finding the disparities at which $\mathbf{V}(\mathbf{p}) = 0$, a reliable semi-dense confident disparity image $\tilde{\mathbf{D}}_{l,r}$ is produced. Finally, $\tilde{\mathbf{D}}_l$ and $\tilde{\mathbf{D}}_r$ are processed by a left-right disparity consistency check (LRDCC) operator to produce $\tilde{\mathbf{D}}$, which is further employed to supervise DCNN training. More details on the DCNN architecture and training phase will be discussed in the next subsections.

B. OptStereo

Given a pair of left and right stereo images $\mathbf{I}_l$ and $\mathbf{I}_r$, our OptStereo is designed to estimate a dense disparity image $\hat{\mathbf{D}}$. Fig. 3 illustrates the overview of our OptStereo, which consists of three stages, *i.e.*, feature extraction, cost volume computation and iterative refinement.

1) Feature Extraction: We use two residual networks [32] that share weights to extract visual features $\mathbf{F}_l$ and $\mathbf{F}_r$ with the size of $H \times W \times C$ from $\mathbf{I}_l$ and $\mathbf{I}_r$, respectively. The spatial size of $\mathbf{F}_l$ and $\mathbf{F}_r$ is 1/8 of the input image resolution, and C is set to be 256. Moreover, the residual networks includes six residual blocks, two at 1/2 resolution, two at 1/4 resolution and two at 1/8 resolution.

2) Cost Volume Computation: In this stage, we compute visual similarity for all possible matching pairs between $\mathbf{F}_l$ and $\mathbf{F}_r$. Specifically, we construct a cost volume $\mathbf{M}^0 \in \mathbb{R}^{H \times W \times W}$ by computing the dot product between all possible matching pairs of feature vectors, which can be formulated as follows:

$$\mathbf{M}^0(i, j, k) = \sum_{h=0}^{C} \mathbf{F}_l(i, j, h) \cdot \mathbf{F}_r(i, k, h). \tag{2}$$

Following [33], we further construct multi-scale cost volumes $\mathbf{M}^1$, $\mathbf{M}^2$ and $\mathbf{M}^3$ by employing average pooling on the last dimension of $\mathbf{M}^0$ respectively, where $\mathbf{M}^k$ has the size of $H \times W \times (W/2^k)$ for $k = 0, 1, 2, 3$. The multi-scale cost volumes $\{\mathbf{M}^0, \mathbf{M}^1, \mathbf{M}^2, \mathbf{M}^3\}$ store information about both large and small disparities, which can be used effectively to update disparity estimations in the iterative refinement stage.

To better utilize the multi-scale cost volumes, we introduce a lookup operation, which can extract values from

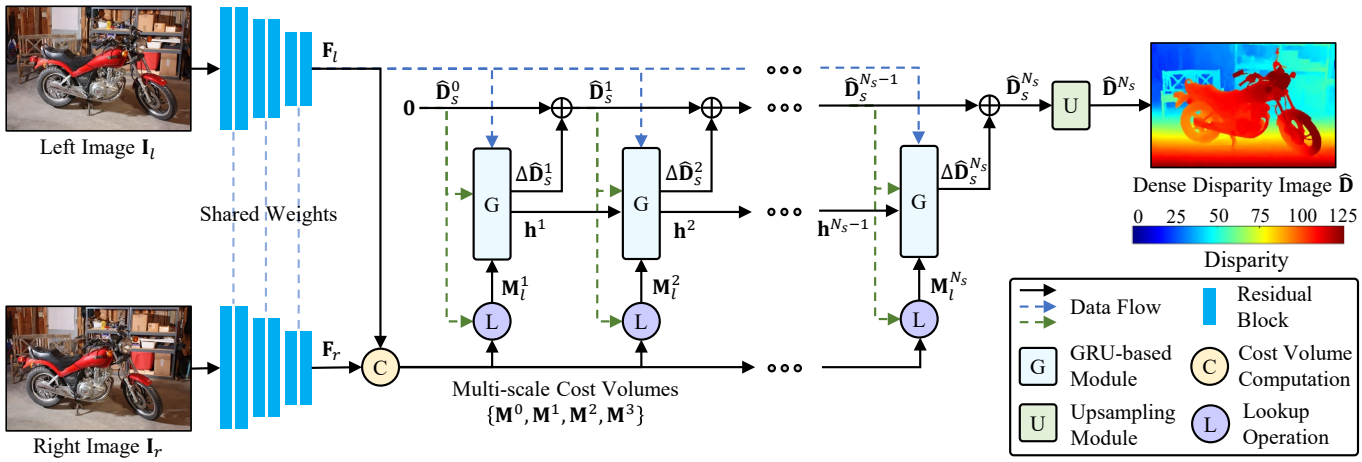

Fig. 3. An illustration of our proposed OptStereo, which first builds multi-scale cost volumes, and then adopts a recurrent unit to iteratively update disparity estimations at high resolution.

$\{\mathbf{M}^0, \mathbf{M}^1, \mathbf{M}^2, \mathbf{M}^3\}$ to construct a local cost volume $\mathbf{M}_l$. Specifically, a dense disparity estimation $\hat{\mathbf{D}}_s$ with the size of $H \times W$ can map a point $\mathbf{p} = (i, j)$ in $\mathbf{F}_l$ to its correspondence $\mathbf{p}' = (i', j') = (i, j - \hat{\mathbf{D}}_s(\mathbf{p}))$ in $\mathbf{F}_r$. We also define a neighbor area around j' as follows:

$$U(j')_d = \{j' + \Delta j | \Delta j \in \mathbb{Z}, |\Delta j| \leq d\}, \tag{3}$$

where d is the constant lookup distance. Then, we can extract values from $\mathbf{M}^k$ by adopting (i, j) as the first two dimensional indexes and every element of $U(j'/2^k)_d$ as the last dimensional index for $k = 0, 1, 2, 3$. Please note that we employ bilinear sampling, since the adopted indexes are real numbers. In addition, we set $d = 4$, which corresponds to a search range of 256 pixels at the input image resolution. The values extracted from the multi-scale cost volumes are then concatenated into a local cost volume $\mathbf{M}_l \in \mathbb{R}^{H \times W \times 36}$, which provides useful visual similarity information around the possible matching pairs indicated by the dense disparity estimation for further refinement.

3) Iterative Refinement: In this stage, we employ a GRU-based module [34] to iteratively update a sequence of dense disparity estimations $\{\hat{\mathbf{D}}_s^1, \ldots, \hat{\mathbf{D}}_s^{N_s}\}$ at the 1/8 resolution with an initialization $\hat{\mathbf{D}}_s^0 = 0$. Specifically, in iteration k, we first use the above-mentioned lookup operation to compute the local cost volume $\mathbf{M}_l^k$ based on the previous dense disparity estimation $\hat{\mathbf{D}}_s^{k-1}$. Then, we denote the concatenation of $\hat{\mathbf{D}}_s^{k-1}$, $\mathbf{M}_l^k$ and $\mathbf{F}_l$ as $\mathbf{x}^k$, and send $\mathbf{x}^k$ to the GRU-based module, which has the following formulation:

$$\begin{aligned}
\mathbf{z}^k &= \sigma\left(\mathrm{Conv}_{3\times3}\left(\left[\mathbf{h}^{k-1}, \mathbf{x}^k\right]\right)\right), \\
\mathbf{r}^k &= \sigma\left(\mathrm{Conv}_{3\times3}\left(\left[\mathbf{h}^{k-1}, \mathbf{x}^k\right]\right)\right), \\
\tilde{\mathbf{h}}^k &= \tanh\left(\mathrm{Conv}_{3\times3}\left(\left[\mathbf{r}^k \odot \mathbf{h}^{k-1}, \mathbf{x}^k\right]\right)\right), \\
\mathbf{h}^k &= \left(1 - \mathbf{z}^k\right) \odot \mathbf{h}^{k-1} + \mathbf{z}^k \odot \tilde{\mathbf{h}}^k,
\end{aligned} \tag{4}$$

where $[\cdot, \cdot]$, σ and $\odot$ denote concatenation, sigmoid function and element-wise multiplication, respectively. The outputted hidden state $\mathbf{h}^k$ is then processed by two convolutional layers to generate the disparity update $\Delta\hat{\mathbf{D}}_s^k$, and the dense disparity estimation $\hat{\mathbf{D}}_s^k$ is updated by $\hat{\mathbf{D}}_s^k = \hat{\mathbf{D}}_s^{k-1} + \Delta\hat{\mathbf{D}}_s^k$. This process iterates until $N_s = 8$ is reached. Note that $\{\hat{\mathbf{D}}_s^1, \ldots, \hat{\mathbf{D}}_s^{N_s}\}$ are at the 1/8 resolution. We then employ an upsampling module that consists of an upsampling layer followed by two convolutional layers to generate the full resolution disparity estimations $\{\hat{\mathbf{D}}^1, \ldots, \hat{\mathbf{D}}^{N_s}\}$. In the inference phase, we take $\hat{\mathbf{D}}^{N_s}$ as the estimated dense disparity image $\hat{\mathbf{D}}$.

The architecture of our OptStereo is inspired by traditional optimization-based approaches. Specifically, the adoption of the GRU-based module mimics the updates of a first-order descent algorithm, and the bounded activations used in (4) also encourage convergence to a fixed point [33]. Moreover, since our OptStereo iteratively updates disparity estimations at high resolution, it does not suffer from the error accumulation problem in the coarse-to-fine paradigm. Our OptStereo can also greatly minimize the trade-off between accuracy and efficiency for stereo matching due to its simple but effective architecture.

C. Loss Function and Data Augmentation

In the training phase, we optimize the parameters of our OptStereo by minimizing a loss function $\mathcal{L}$, which consists of three terms:

$$\mathcal{L} = \mathcal{L}_P + \lambda_1 \cdot \mathcal{L}_R + \lambda_2 \cdot \mathcal{L}_S, \tag{5}$$

where $\mathcal{L}_P$ denotes the PVM guiding loss; $\mathcal{L}_R$ denotes the reconstruction loss; $\mathcal{L}_S$ denotes the smoothing loss; and λ_1 and λ_2 are two hyper-parameters to weight the contributions of the three above-mentioned loss terms.

The $\mathcal{L}_P$ term takes the reliable semi-dense disparity images generated by our PVM as the supervision for training. Specifically, we adopt the Huber loss function $l(\cdot)$ for the $\mathcal{L}_P$ term, which is defined as:

$$\begin{aligned}
\mathcal{L}_P &= \frac{1}{\sum_{i=1}^{N_s} \gamma^{N_s - i}} \sum_{i=1}^{N_s} \frac{\gamma^{N_s - i}}{N_{\tilde{\mathbf{D}}}} \sum_{\mathbf{p} \in \tilde{\mathbf{D}}} l(|\tilde{\mathbf{D}}(\mathbf{p}) - \hat{\mathbf{D}}^i(\mathbf{p})|), \\
l(x) &= \begin{cases} x - 0.5, & x \geq 1 \\ x^2/2, & x < 1 \end{cases},
\end{aligned} \tag{6}$$

where $\tilde{\mathbf{D}}$ denotes the reliable semi-dense disparity image generated by our PVM; $\hat{\mathbf{D}}^i$ denotes the dense disparity image estimated by our OptStereo; $N_{\tilde{\mathbf{D}}}$ is the number of observed pixels in $\tilde{\mathbf{D}}$; and γ is set to be 0.8 in our experiments. Compared with the $L2$ loss, the Huber loss function $l(\cdot)$ has lower sensitivity to outliers and presents more robust performance at discontinuous areas [35].

Since the $\mathcal{L}_P$ term only considers the sparse pixels, we include the $\mathcal{L}_R$ term to add constraints on the densely predicted pixels. Inspired by [36], given a pair of stereo images, the left image $\mathbf{I}_l$ can be reconstructed from the right image $\mathbf{I}_r$ based on the estimated disparity image $\hat{\mathbf{D}}$. To make the reconstruction process differentiable, we employ a bilinear sampler in this process. Finally, the $\mathcal{L}_R$ term is defined as a combination of a single-scale SSIM term [37] and an $L1$ norm term:

$$\mathcal{L}_R = \frac{1}{N_{\hat{\mathbf{I}}_l}} \sum_{\mathbf{p}\in\hat{\mathbf{I}}_l} \alpha \frac{1-\text{SSIM}\left(\mathbf{I}_l(\mathbf{p}), \hat{\mathbf{I}}_l(\mathbf{p})\right)}{2} + (1-\alpha)\left\|\mathbf{I}_l(\mathbf{p}) - \hat{\mathbf{I}}_l(\mathbf{p})\right\|_1, \tag{7}$$

where $\hat{\mathbf{I}}_l$ is the reconstructed image from $\mathbf{I}_r$ according to $\hat{\mathbf{D}}$; and $N_{\hat{\mathbf{I}}_l}$ is the number of observed pixels in $\hat{\mathbf{I}}_l$. Similar to [36], we use a simplified SSIM with a 3×3 block filter and set $\alpha = 0.85$ in our experiments.

Inspired by [38], we further add the $\mathcal{L}_S$ term to smooth the disparity predictions. Since the disparity values at the place where the image pixel intensity changes greatly usually vary significantly, we define the $\mathcal{L}_S$ term as weighting the disparity gradients ($\partial\hat{\mathbf{D}}$) with an edge-aware term using the image gradients ($\partial\mathbf{I}_l$):

$$\mathcal{L}_S = \frac{1}{N_{\hat{\mathbf{D}}}} \sum_{\mathbf{p}\in\hat{\mathbf{D}}} |\partial_x\hat{\mathbf{D}}(\mathbf{p})|e^{-\|\partial_x\mathbf{I}_l(\mathbf{p})\|_1} + |\partial_y\hat{\mathbf{D}}(\mathbf{p})|e^{-\|\partial_y\mathbf{I}_l(\mathbf{p})\|_1}. \tag{8}$$

Furthermore, in traditional stereo matching approaches, the LRDCC is usually performed to refine the estimated disparities, as the occluded areas are only visible in one image. Inspired by the LRDCC, we flip the left and right stereo images in the left-right direction, respectively, as shown in Fig. 1. The flipped left and right images are considered as a pair of new right and left images, respectively. This process can augment the training data to improve the stereo matching performance of DCNNs.

IV. Experimental Results and Discussion

A. Datasets and Implementation Details

We use three stereo datasets in our experiments:

- The KITTI Stereo 2012 benchmark [17]: This dataset contains 194 training stereo image pairs with sparse ground-truth disparity images and another 195 testing stereo image pairs without ground-truth disparity images.
- The KITTI Stereo 2015 benchmark [18]: This dataset contains 200 training stereo image pairs with sparse ground-truth disparity images and another 200 testing stereo image pairs without ground-truth disparity images.
- Our HKUST-Drive dataset: We publish a large-scale synthetic dataset, named the HKUST-Drive dataset. This dataset is created using the CARLA simulator [39]. It is collected in six different scenarios under different illumination and weather conditions, *e.g.,* clear, rainy, daytime and sunset. There are a total 11568 pairs of stereo images with corresponding dense sub-pixel disparity ground truth. We split it into a training set (6940 image pairs), a validation set (2314 image pairs) and a testing set (2314 image pairs). Different from the KITTI stereo datasets [17], [18], our HKUST-Drive dataset can effectively evaluate the generalization ability of stereo matching approaches across different weather and illumination conditions.

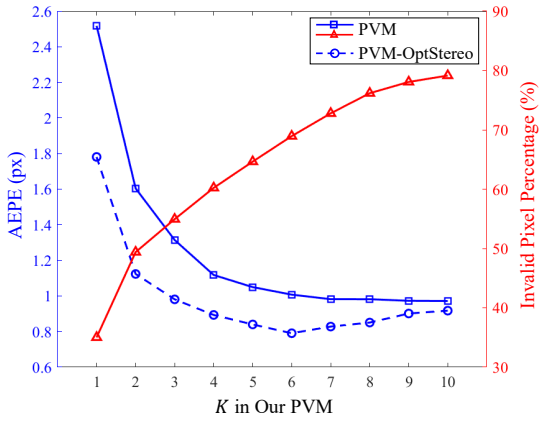

Fig. 4. Experimental results of the proposed approach with different variants of our PVM on the HKUST-Drive dataset.

During the training phase, we use the Adam optimizer [40] and adopt an initial learning rate of 10^{-4}. The model is trained on two NVIDIA GeForce RTX 2080 Ti graphics cards until it converges. We also adopt an existing self-supervision scheme [41] to improve the stereo matching performance on the challenging areas, such as the occluded areas. Moreover, we use two commonly used metrics for evaluation: 1) the average end-point error (AEPE) that measures the average difference between the disparity estimations and ground-truth labels and 2) the percentage of pixels (F1) with absolute disparity error higher than 3 pixels [17].

In our experiments, we first conduct ablation studies on our HKUST-Drive dataset in Section IV-B to 1) select the best architecture of our PVM and OptStereo as well as the best hyper-parameters for the loss function; and 2) demonstrate the effectiveness of our data augmentation technique. Then, we denote our best self-supervised implementation as *PVStereo*, and compare it with the state-of-the-art self-supervised approaches on our HKUST-Drive dataset. We also compare our OptStereo with the state-of-the-art supervised approaches. The experimental results are shown in Section IV-C. After that, we evaluate our PVStereo and OptStereo on the popular KITTI Stereo 2012 and 2015 benchmarks [17], [18], as presented in Section IV-D.

B. Ablation Study

We first explore the best architecture of our PVM. Fig. 4 compares the performance of our PVM and the corresponding

TABLE I
EXPERIMENTAL RESULTS OF THE PROPOSED APPROACH WITH DIFFERENT VARIANTS OF OUR OPTSTEREO ON THE HKUST-DRIVE DATASET. THE ADOPTED VARIANT IS BOLDED

No.	Multi-scale Cost Volumes	Lookup Distance d	AEPE (px)
(a)	–	4	1.02
(b)	✓	1	1.16
(c)	✓	2	0.93
(d)	✓	4	**0.79**

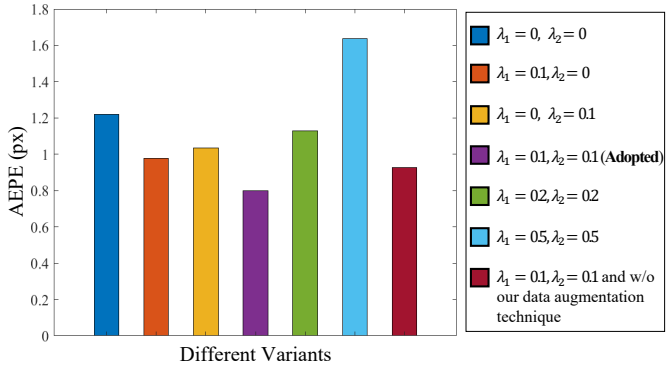

Fig. 5. Experimental results of the proposed approach with different variants of our loss function and data augmentation technique on the HKUST-Drive dataset.

PVM-OptStereo with respect to different K. We can observe that with K increasing, the AEPE of the PVM decreases while the invalid pixel percentage increases. Correspondingly, the AEPE of our PVM-OptStereo first decreases but then increases. Therefore, we minimize the trade-off between accuracy and density, and set K to 6 in the rest of our experiments, where the corresponding PVM-OptStereo can achieve the best performance. Additionally, we can also observe that all the variants of our PVM-OptStereo outperform the corresponding PVM. We believe that our designed loss function and proposed data augmentation technique can provide effective fine-tuning for the predicted dense disparity images, thus making them outperform the corresponding semi-dense disparity images used for training.

Table I compares the performance of the proposed approach with different variants of our OptStereo. The comparison between (a) and (d) demonstrates the effectiveness of the adopted multi-scale cost volumes. From (b)–(d), we can observe that with the lookup distance d increasing, the performance of our approach improves. Considering the balance between accuracy and efficiency, we adopt (d) of Table I in the rest of our experiments.

In addition, we test different combinations of λ_1 and λ_2 in the loss function, and some of the experimental results are presented in Fig. 5. We can observe that the proper introduction of $\mathcal{L}_R$ and $\mathcal{L}_S$ can improve the performance effectively. We analyze that $\mathcal{L}_R$ and $\mathcal{L}_S$ can perform effective supervision for the pixels that are invalid in our PVM, and thus, can benefit the overall performance of dense stereo matching. However,

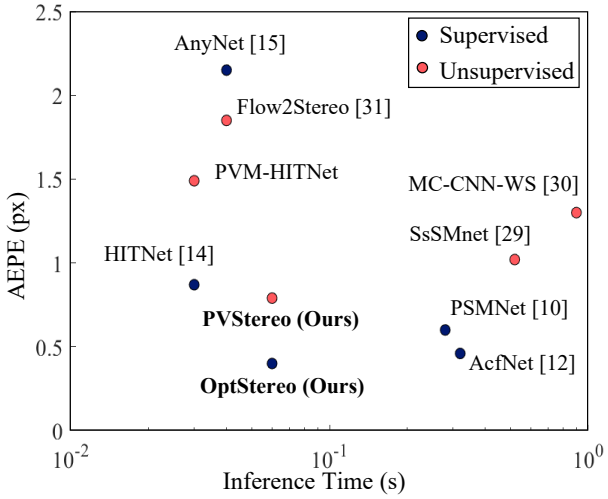

Fig. 6. Performance comparison of our OptStereo and PVStereo with state-of-the-art supervised and unsupervised approaches on the HKUST-Drive dataset.

λ_1 and λ_2 must be kept low to not overcome the contribution of $\mathcal{L}_P$. Based on Fig. 5, we set $\lambda_1 = 0.1$ and $\lambda_2 = 0.1$ in the rest of our experiments. Furthermore, Fig. 5 demonstrates that our data augmentation technique can effectively improve the stereo matching performance. We analyze that the proposed technique can leverage the relationship among stereo images to perform effective training data augmentation, and thus, can benefit the overall performance of dense stereo matching.

C. Performance Comparison on Our HKUST-Drive Dataset

As previously mentioned, we denote our best self-supervised implementation as *PVStereo*, and compare it with the state-of-the-art self-supervised approaches on our HKUST-Drive dataset. We also train our OptStereo in a supervised manner, and compare it with the state-of-the-art supervised approaches. In addition, we implement our PVM in existing HITNet [14], and denote it as PVM-HITNet. For the other baseline supervised and unsupervised approaches, we follow the hyperparameter setups as reported in their papers.

The quantitative results are presented in Fig. 6. We can clearly observe that our PVStereo significantly outperforms all other state-of-the-art self-supervised approaches. Moreover, our PVStereo can achieve a great trade-off between accuracy and efficiency for stereo matching. It is also evident that PVM-HITNet can present competitive performance for self-supervised stereo matching, which verifies the effectiveness of our PVM. Furthermore, our PVStereo can even achieve competitive performance compared to supervised approaches such as HITNet [14], which demonstrates the effectiveness of our proposed self-supervised architecture. Please note that our OptStereo achieves a better performance than state-of-the-art supervised approaches, which verifies the superiority of the proposed DCNN architecture over existing architectures that rely on 3D convolutions or coarse-to-fine paradigms. Fig. 7 illustrates an example of the qualitative results. We can see that our PVStereo can generate more accurate disparity estimations. Since our HKUST-Drive dataset covers different

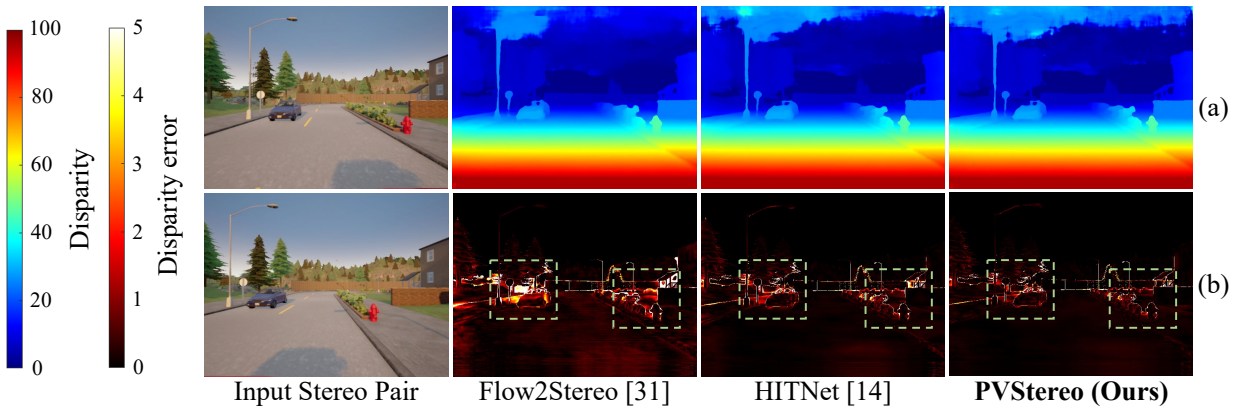

Fig. 7. An example on our HKUST-Drive dataset, where rows (a) and (b) show the disparity estimations and the corresponding disparity error maps, respectively. Significantly improved regions are marked with green dashed boxes.

TABLE II
EVALUATION RESULTS ON THE KITTI STEREO 2012[1] [17] AND KITTI STEREO 2015[2] [18] BENCHMARKS. "NOC" AND "ALL" REPRESENT THE F1 (%) FOR NON-OCCLUDED PIXELS AND ALL PIXELS, RESPECTIVELY. "S" DENOTES SUPERVISED APPROACHES, AND PSGM [25] IS A TRADITIONAL APPROACH. BEST RESULTS FOR SUPERVISED AND UNSUPERVISED APPROACHES ARE BOTH BOLDED

Approach	S	KITTI 2012		KITTI 2015		Runtime (s)
		Noc	All	Noc	All	
PSMNet [10]	✓	1.49	1.89	2.14	2.32	0.41
GwcNet-g [11]	✓	1.37	1.70	1.92	2.11	0.32
AcfNet [12]	✓	**1.17**	**1.54**	1.72	1.89	0.48
OptStereo (Ours)	✓	1.20	1.61	**1.36**	**1.82**	**0.10**
pSGM [25]	–	4.68	6.13	5.17	5.97	7.77
Flow2Stereo [31]	–	4.58	5.11	6.29	6.61	**0.05**
MC-CNN-WS [30]	–	3.02	4.45	4.11	4.97	1.35
SsSMnet [29]	–	2.30	3.00	3.06	3.40	0.80
PVStereo (Ours)	–	**1.98**	**2.47**	**2.69**	**2.99**	0.10

scenarios under different illumination and weather conditions, these results strongly demonstrate the great generalization ability of our OptStereo and PVStereo across different scenarios as well as different weather and illumination conditions.

D. Evaluation Results on the KITTI Stereo Benchmarks

We submit the results achieved by our OptStereo and PVStereo to KITTI Stereo benchmarks [17], [18], and the quantitative results are presented in Table II. It is evident that our OptStereo achieves competitive performance compared to state-of-the-art supervised stereo matching approaches, which demonstrates the effectiveness of the proposed DCNN architecture. Moreover, our PVStereo outperforms all other state-of-the-art self-supervised stereo matching approaches with a great trade-off between accuracy and efficiency, which verifies the superiority of our self-supervised stereo matching approach. Furthermore, Fig. 8 illustrates some examples on the KITTI Stereo benchmarks, where we can see that our OptStereo and PVStereo can yield robust and accurate disparity estimations.

[1]cvlibs.net/datasets/kitti/eval_stereo_flow.php?benchmark=stereo
[2]cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=stereo

V. CONCLUSIONS

This paper presented PVStereo, a novel self-supervised approach for end-to-end stereo matching, which consists of a PVM and a novel DCNN architecture, referred to as OptStereo. Specifically, our OptStereo first builds multi-scale cost volumes, and then adopts a recurrent unit to iteratively update disparity estimations at high resolution, which can not only avoid the error accumulation problem in coarse-to-fine paradigms, but also can achieve a great trade-off between accuracy and efficiency due to its simple but effective architecture. Moreover, our PVM can generate reliable semi-dense disparity images, which can be employed to supervise the training of OptStereo. Furthermore, we publish a large-scale synthetic stereo dataset, named the HKUST-Drive dataset, collected under different illumination and weather conditions for research purposes. Extensive experiments on the popular KITTI Stereo benchmarks and our HKUST-Drive dataset demonstrate the effectiveness and efficiency of our PVStereo, which greatly outperforms all other state-of-the-art self-supervised stereo matching approaches. We believe that our PVStereo can be employed in many robotics applications, such as freespace detection, to improve their performance. It is also promising to employ the proposed architecture in other self-supervised tasks, such as self-supervised optical flow estimation.

REFERENCES

[1] N. Qian, "Binocular disparity and the perception of depth," *Neuron*, vol. 18, no. 3, pp. 359–368, 1997.
[2] R. Fan, L. Wang, M. J. Bocus, and I. Pitas, "Computer stereo vision for autonomous driving," *CoRR*, 2020.
[3] R. Fan, H. Wang, P. Cai, and M. Liu, "Sne-roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection," in *Eur. Conf. Comput. Vision (ECCV)*. Springer, 2020, pp. 340–356.
[4] H. Wang, R. Fan, Y. Sun, and M. Liu, "Applying surface normal information in drivable area and road anomaly detection for ground mobile robots," in *IEEE/RSJ Int. Conf. on Intell. Robots Syst.*, 2020.
[5] R. Fan, H. Wang, P. Cai, J. Wu, M. J. Bocus, L. Qiao, and M. Liu, "Learning collision-free space detection from stereo images: Homography matrix brings better data augmentation," *IEEE/ASME Trans. Mechatronics*, 2021.
[6] H. Wang, Y. Sun, and M. Liu, "Self-supervised drivable area and road anomaly segmentation using rgb-d data for robotic wheelchairs," *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 4386–4393, 2019.

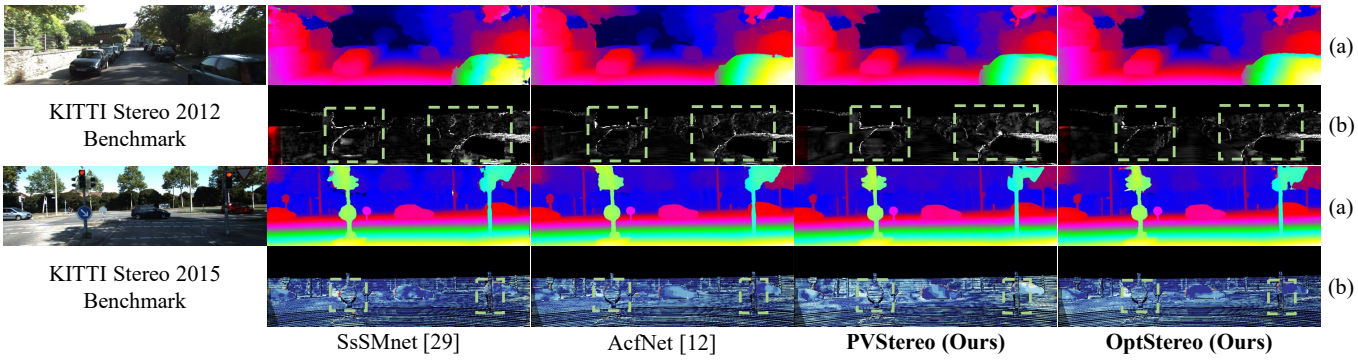

Fig. 8. Examples on the KITTI Stereo benchmarks [17], [18], where rows (a) and (b) show the disparity estimations and the corresponding disparity error maps, respectively. Significantly improved regions are highlighted with green dashed boxes.

[7] R. Fan, H. Wang, M. J. Bocus, and M. Liu, "We learn better road pothole detection: from attention aggregation to adversarial domain adaptation," in *Eur. Conf. Comput. Vision (ECCV)*. Springer, 2020, pp. 285–300.

[8] H. Wang, R. Fan, Y. Sun, and M. Liu, "Dynamic fusion module evolves drivable area and road anomaly detection: A benchmark and algorithms," *IEEE Trans. Cybern.*, 2021.

[9] V. Q. Dinh, F. Munir, A. M. Sheri, and M. Jeon, "Disparity estimation using stereo images with different focal lengths," *IEEE Trans. Intell. Transp. Syst.*, 2019.

[10] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2018, pp. 5410–5418.

[11] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2019, pp. 3273–3282.

[12] Y. Zhang, Y. Chen, X. Bai, S. Yu, K. Yu, Z. Li, and K. Yang, "Adaptive unimodal cost volume filtering for deep stereo matching." in *AAAI*, 2020, pp. 12 926–12 934.

[13] N. Smolyanskiy, A. Kamenev, and S. Birchfield, "On the importance of stereo for accurate depth estimation: An efficient semi-supervised deep neural network approach," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshop (CVPRW)*, 2018, pp. 1007–1015.

[14] V. Tankovich, C. Häne, S. Fanello, Y. Zhang, S. Izadi, and S. Bouaziz, "Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching," *arXiv preprint arXiv:2007.12140*, 2020.

[15] Y. Wang, Z. Lai, G. Huang, B. H. Wang, L. Van Der Maaten, M. Campbell, and K. Q. Weinberger, "Anytime stereo image depth estimation on mobile devices," in *2019 Int. Conf. Robot. Automat. (ICRA)*. IEEE, 2019, pp. 5893–5900.

[16] K. Yee and A. Chakrabarti, "Fast deep stereo with 2d convolutional processing of cost signatures," in *IEEE Winter Conf. Appl. Comput. Vision (WACV)*, 2020, pp. 183–191.

[17] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*. IEEE, 2012, pp. 3354–3361.

[18] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2015, pp. 3061–3070.

[19] R. Fan, X. Ai, and N. Dahnoun, "Road surface 3d reconstruction based on dense subpixel disparity map estimation," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3025–3035, 2018.

[20] H. Hirschmuller and D. Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1582–1599, 2008.

[21] B. Tippetts, D. J. Lee, K. Lillywhite, and J. Archibald, "Review of stereo vision algorithms and their suitability for resource-limited systems," *J. Real-Time Image Process.*, vol. 11, no. 1, pp. 5–25, 2016.

[22] R. Fan, "Real-time computer stereo vision for automotive applications," Ph.D. dissertation, Dept. of Electrical & Electronic Eng, University of Bristol, 2018.

[23] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, 2001.

[24] A. T. Ihler, W. F. John III, and A. S. Willsky, "Loopy belief propagation: Convergence and effects of message errors," *J. Mach. Learn. Res.*, vol. 6, no. May, pp. 905–936, 2005.

[25] Y. Lee, M.-G. Park, Y. Hwang, Y. Shin, and C.-M. Kyung, "Memory-efficient parametric semiglobal matching," *IEEE Signal Process. Lett.*, vol. 25, no. 2, pp. 194–198, 2017.

[26] R. Fan, U. Ozgunalp, Y. Wang, M. Liu, and I. Pitas, "Rethinking road surface 3d reconstruction and pothole detection: From perspective transformation to disparity map segmentation," *IEEE Trans. Cybern.*, 2020.

[27] J. Žbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *The J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2287–2318, 2016.

[28] A. Seki and M. Pollefeys, "Sgm-nets: Semi-global matching with neural networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2017, pp. 231–240.

[29] Y. Zhong, Y. Dai, and H. Li, "Self-supervised learning for stereo matching with self-improving ability," *arXiv preprint arXiv:1709.00930*, 2017.

[30] S. Tulyakov, A. Ivanov, and F. Fleuret, "Weakly supervised learning of deep metrics for stereo reconstruction," in *Proc. IEEE Inter. Conf. Comput. Vision (ICCV)*, 2017, pp. 1339–1348.

[31] P. Liu, I. King, M. Lyu, and J. Xu, "Flow2stereo: Effective self-supervised learning of optical flow and stereo matching," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2020.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[33] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Eur. Conf. Comput. Vision (ECCV)*. Springer, 2020, pp. 402–419.

[34] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[35] P. J. Huber, "Robust estimation of a location parameter," in *Breakthroughs Statist.* Springer, 1992, pp. 492–518.

[36] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2017, pp. 270–279.

[37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[38] P. Heise, S. Klose, B. Jensen, and A. Knoll, "Pm-huber: Patchmatch with huber regularization for stereo matching," in *Proc. IEEE Inter. Conf. Comput. Vision (ICCV)*, 2013, pp. 2360–2367.

[39] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. 1st Annu. Conf. Robot Learn. (CoRL)*, ser. Proc. Mach. Learn. Res., S. Levine, V. Vanhoucke, and K. Goldberg, Eds., vol. 78. PMLR, 13–15 Nov 2017, pp. 1–16.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[41] L. Liu, J. Zhang, R. He, Y. Liu, Y. Wang, Y. Tai, D. Luo, C. Wang, J. Li, and F. Huang, "Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2020, pp. 6489–6498.